

Research on Multi Source Data Driven Modeling of Individual Investment Decision Behavior and Machine Learning Prediction Mechanism

Wenjia Zhang¹, Yu Zhang², Yexin Zhang^{3*}

¹College of Business and Economics, Australian National University, Canberra, Australia

²College of Management, Beijing Union University, Beijing, China

³Monash Business School, Monash University, Melbourne, Australia

*Corresponding Author. Email: rara481846778@gmail.com

Abstract. This study develops a multi-source data-driven framework for modeling individual investment decision behavior and predicting investor actions through machine learning. The research focuses on how trading history, portfolio structure, market exposure, information attention, and user profile variables jointly shape next-day buy, sell, and hold decisions. The dataset is constructed from brokerage transaction records, investor account profiles, daily stock-market quotation data, and financial-news interaction logs, with all records aligned at the user-day level. A fixed feature system is built through rolling-window trading indicators, portfolio-risk indicators, market-state variables, attention features, and demographic-account variables. The prediction mechanism is implemented through an XGBoost multi-class classifier, supported by standardized feature processing, stratified sample splitting, probability output generation, and SHAP-based interpretation. The results show that the model achieves stable prediction performance on the test set, with macro-F1 of 0.759 ± 0.008 and multi-class AUC of 0.864 ± 0.004 . Trading frequency, unrealized return, cash proportion, market volatility, and news-click intensity are the strongest behavioral predictors.

Keywords: Individual investment behavior, multi-source data, machine learning prediction, investor decision modeling, behavioral finance

1. Introduction

Individual investment behavior is increasingly recorded as digital traces generated through trading accounts, portfolio movements, market-information access, and mobile-platform interactions. These traces make it possible to examine investor decisions not only through survey-based attitudes but also through repeated behavioral records at the observation level. Recent financial research has emphasized that big data can reshape empirical finance because trading, attention, and platform-use records provide richer signals than traditional aggregate market data [1]. Studies on smartphone-based investing further show that digital access channels can alter trading intensity and risk-taking behavior [2]. Evidence from crisis-period trading indicates that individual investors' reactions

depend on market stress, prior account position, and perceived buying opportunities [3]. Machine-learning studies have demonstrated the value of clustering techniques for identifying heterogeneous risk-behavior patterns among investors [4]. Research has also shown that online social networks increasingly influence how retail investors acquire information and make investment decisions [5]. In addition, robo-advisory systems can affect investor behavior through digital guidance and platform design mechanisms [6]. Studies on financial literacy, self-control, and investment decision-making indicate that personal characteristics remain important determinants of investment outcomes [7]. More recent evidence suggests that machine-learning methods can effectively predict investor switching behavior [8], while deep-learning and clustering approaches reveal substantial heterogeneity in individual trading patterns [9]. Research on robo-advisers and investor behavior further confirms the growing importance of technology-driven investment environments [10]. The framework combines brokerage records, account profiles, quotation data, and financial-news click logs into one user-day dataset. It then builds fixed feature groups, trains an XGBoost multi-class classifier, evaluates prediction performance, and explains behavioral mechanisms through feature contribution analysis.

2. Literature review

Individual investment decision behavior has been widely discussed in behavioral finance. With digital trading platforms, investor actions can now be observed through transaction records, account operations, and online information traces. Existing research suggests that investment decisions are influenced by expected returns, prior trading experience, portfolio performance, liquidity conditions, and market volatility [1, 2]. During uncertain market periods, investors respond differently according to account position, perceived risk, and available opportunities [3]. Financial literacy, self-control, and behavioral traits also affect investment choices and financial well-being [7]. These findings indicate that investor behavior is dynamic and heterogeneous, requiring data-driven models that can capture individual differences in repeated decision processes [9]. The growing availability of financial data has made multi-source integration increasingly important in investor behavior research. Market quotation data provide external information on price changes, volatility, and trading volume. Online information activities, including financial-news browsing and platform interactions, further reflect investor attention and information acquisition. Previous studies show that combining trading behavior with attention-related indicators improves the understanding of investor decision mechanisms [1, 2]. Research on social networks and robo-advisory services also suggests that digital interactions influence information processing and portfolio adjustment [5, 6].

3. Experimental methodology

3.1. Data sources and observation unit construction

The dataset is built from four fixed sources: brokerage transaction records, investor account profile tables, daily stock-market quotation data, and financial-news click logs. Transaction records include investor ID, stock code, order direction, executed volume, executed value, transaction timestamp, commission, and realized gain or loss. Account profile tables include investor age group, account opening length, risk preference score, account capital level, and product-permission category. Market quotation data include daily open price, close price, intraday high-low range, turnover rate, trading volume change, market-index return, and sector-index return. Financial-news click logs include firm-news clicks, sector-news clicks, market-page visits, reading duration, and last-news-

click interval. All records are stored and extracted through MySQL 8.0. Investor IDs are encrypted before analysis. The observation period contains 365 consecutive trading days. Only investors with continuous account existence and valid portfolio records during the period are retained. The final sample includes 12,000 individual investors and 180,000 valid user-day observations after removing inactive accounts, incomplete quotation matches, and abnormal negative-balance records. Each observation represents one investor on one trading day. The dependent variable is the investor's next-day primary action. If net purchase value is positive and exceeds the transaction threshold, the label is buy. If net sale value is positive and exceeds the threshold, the label is sell. If neither condition is met, the label is hold.

3.2. Feature engineering and behavioral label definition

Feature engineering is completed in Python 3.11 using pandas, NumPy, and scikit-learn. The feature system is divided into five groups. Trading-history features include trade frequency in the past 5 days, turnover ratio, realized gain ratio, average holding period, repeated trade indicator, and previous-day action type. Portfolio-structure features include cash proportion, stock concentration, unrealized return, sector diversification index, position volatility, and loss-position share. Market-exposure features include market-index return, stock return, sector return, intraday volatility, price momentum, volume change, and market drawdown. Information-attention features include financial-news clicks, sector-news click share, average reading duration, market-page time, and last-news-click interval. Investor-profile features include age group, account opening length, risk preference score, and capital level. The user-day feature vector is constructed through rolling-window aggregation:

$$X_{i,t} = \left[\frac{1}{w} \sum_{\tau=t-w}^{t-1} T_{i,\tau}, \frac{V_{i,t-1}^{\text{stock}}}{V_{i,t-1}^{\text{asset}}}, \frac{\text{Cash}_{i,t-1}}{\text{Asset}_{i,t-1}}, \sum_{s=1}^S q_{i,s,t-1}^2, R_{m,t-1}, \sigma_{p,i,t-w}, \log(1 + \text{NewsClick}_{i,t-1}), \text{Profile}_i \right] \quad (1)$$

where $X_{i,t}$ is the feature vector of investor i on day t , $T_{i,\tau}$ is historical trading activity, V^{stock} is stock position value, V^{asset} is total asset value, Cash is cash balance, $q_{i,s,t-1}$ is sector position share, R_m is market return, σ_p is portfolio volatility, and Profile_i represents stable account-profile variables. The target variable $Y_{i,t+1}$ contains three classes: buy, sell, and hold. Feature values are calculated using only information available before the prediction date.

3.3. Prediction model and evaluation configuration

The prediction model is a fixed XGBoost multi-class classifier implemented in XGBoost 2.0. The hyperparameters are set as `max_depth = 6`, `learning_rate = 0.05`, `n_estimators = 300`, `subsample = 0.8`, `colsample_bytree = 0.8`, `min_child_weight = 3`, `reg_lambda = 1.2`, and `random_state = 42`. Continuous variables are standardized with `StandardScaler`, and categorical variables are transformed through one-hot encoding. The model outputs three probability values for each user-day observation. Evaluation metrics include accuracy, macro-F1, multi-class AUC, and log loss. SHAP `TreeExplainer` is used to identify global feature contributions and segment-specific behavioral mechanisms.

4. Experimental procedure

4.1. Data cleaning, alignment, and sample splitting

The raw transaction table is first cleaned by deleting duplicate order records, canceled trades, missing execution prices, invalid investor IDs, and trades outside the official trading calendar. Rolling features are calculated with 5-day, 10-day, and 20-day windows, but only the fixed 5-day features are retained in the main model because validation performance is most stable under this window. The final analytic table is divided into 70% training set, 15% validation set, and 15% test set. The split is stratified by buy, sell, and hold labels, and all records from the same investor are kept in chronological order to preserve behavioral continuity.

4.2. Model training, validation, and prediction output generation

The XGBoost classifier is trained on the training set using the fixed feature matrix and three-class label vector. The validation set is used to monitor macro-F1, log loss, and class-wise recall. Early stopping is applied if validation log loss does not improve for 30 consecutive rounds, while the maximum number of boosting rounds remains fixed at 300. The selected model is then frozen and evaluated only once on the independent test set. The model generates probability outputs $P(\text{buy})$, $P(\text{sell})$, and $P(\text{hold})$ for every test observation. The final prediction class is assigned by the maximum probability rule. If the highest probability is below 0.45, the model still assigns the largest class, but the record is marked as low-confidence in the probability table. This makes the prediction output usable for later behavioral monitoring. For each test observation, the output table stores investor ID, date, true label, predicted label, three class probabilities, confidence score, and major contributing features.

4.3. Behavioral interpretation and segment-based prediction analysis

Behavioral interpretation is conducted with SHAP on the full test set. Global results rank features by mean absolute SHAP value, while class-specific results examine buy, sell, and hold predictions separately. Trade frequency reflects the effect of recent repeated trading on active responses, while unrealized return captures how gain or loss positions influence selling or holding decisions. News-click intensity is interpreted with last-news-click interval to distinguish general attention from immediate pre-trade attention. Segment analysis further compares high-frequency, low-frequency, high-capital, and low-capital investors using the same trained model.

5. Experimental results

5.1. Descriptive statistics of multi-source investor data and behavioral label distribution

The final dataset contains 180,000 user-day observations from 12,000 investors. The average 5-day trade frequency is 1.842 ± 1.317 , the average cash proportion is 0.286 ± 0.174 , and the mean unrealized return is 0.037 ± 0.119 (see Figure 1). News-click intensity shows strong dispersion, with 3.684 ± 4.912 clicks per user-day. The behavioral labels show that hold is the dominant class, but buy and sell actions remain sufficiently represented for supervised classification.

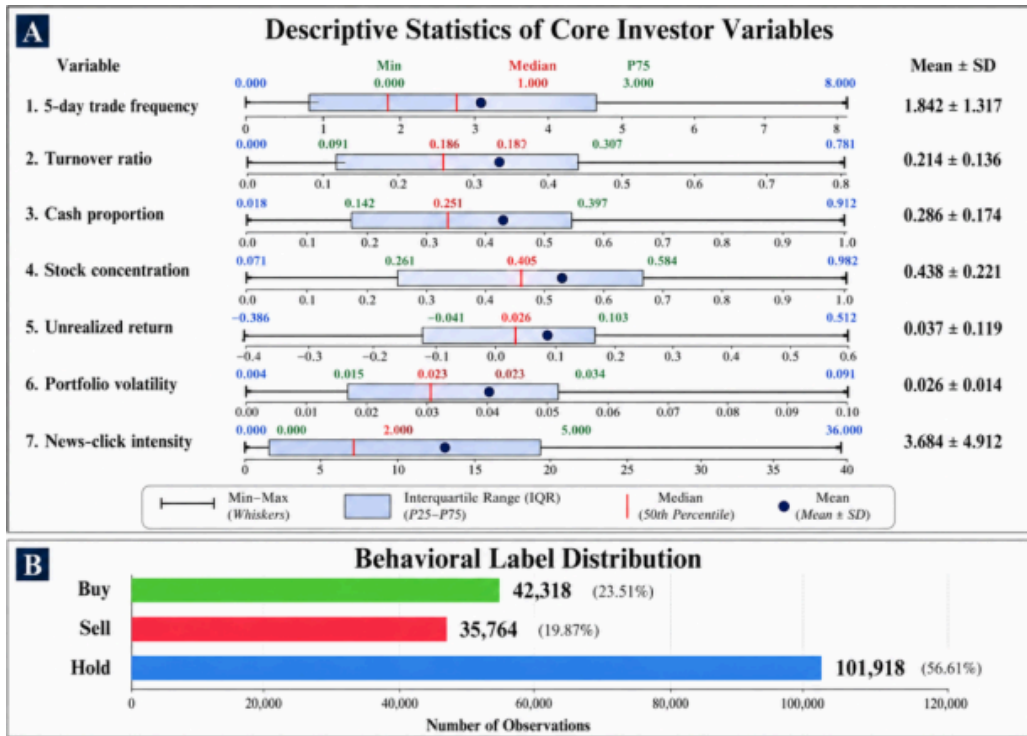


Figure 1. Descriptive statistics and behavioral label distribution

5.2. Prediction performance of buy, sell, and hold decisions under the XGBoost mechanism

The XGBoost model shows stable test performance. Overall accuracy reaches 0.782 ± 0.006 , macro-F1 is 0.759 ± 0.008 , multi-class AUC is 0.864 ± 0.004 , and log loss is 0.486 ± 0.013 . The hold class has the highest recall, while buy and sell predictions are more sensitive to trading-frequency and portfolio-pressure features. Compared with a multinomial logistic benchmark, the XGBoost model improves macro-F1 by $0.094 \pm 0.011^*$ and reduces log loss by $0.137 \pm 0.019^*$ (see Table 1).

Table 1. Prediction performance and segment-level results

Model or segment	Accuracy	Macro-F1	Multi-class AUC	Log loss
Multinomial logit	0.688±0.009	0.665±0.010	0.742±0.007	0.623±0.018
Random forest	0.741±0.007	0.716±0.009	0.821±0.006	0.541±0.016
XGBoost full model	0.782±0.006*	0.759±0.008*	0.864±0.004*	0.486±0.013*
High-frequency traders	0.806±0.007*	0.781±0.009*	0.881±0.005*	0.451±0.014*
Low-frequency traders	0.731±0.010	0.702±0.012	0.812±0.008	0.559±0.019
High-capital investors	0.794±0.008*	0.768±0.009*	0.873±0.005*	0.469±0.015*
Low-capital investors	0.746±0.009	0.721±0.011	0.829±0.007	0.532±0.017

5.3. Feature contribution ranking and segment-level behavioral prediction differences

SHAP results show that 5-day trade frequency has the highest mean absolute contribution value, reaching $0.146 \pm 0.021^*$. Unrealized return follows with $0.118 \pm 0.018^*$, while cash proportion reaches $0.107 \pm 0.016^*$. Market volatility and news-click intensity contribute $0.091 \pm 0.014^*$ and $0.084 \pm 0.013^*$, respectively. High-frequency traders are mainly predicted by recent trading rhythm,

while low-frequency traders depend more on cash reserve and unrealized return. High-capital investors show stronger sensitivity to portfolio volatility. Low-capital investors show higher dependence on financial-news clicks, indicating that information attention plays a larger role when account size is smaller.

6. Conclusion

This study constructs a multi-source data-driven framework for predicting next-day buy, sell, and hold decisions. By integrating transaction records, account profiles, quotation data, and financial-news logs, the model captures investor behavior at the user-day level. The XGBoost classifier shows strong performance, with macro-F1 of 0.759 ± 0.008 and AUC of 0.864 ± 0.004 . The results confirm that investment decisions are jointly shaped by trading history, portfolio pressure, market state, and information attention.

Author contribution

Wenjia Zhang and Yu Zhang contributed equally to this paper.

References

- [1] Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., & Huang, M. (2021). Towards emotional support dialog systems. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 3469–3483.
- [2] Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digital Medicine*, 6, 236.
- [3] Cheng, Y., Zhang, W., Liu, T., He, S., & Zhao, D. (2022). Improving multi-turn emotional support dialogue generation with lookahead strategy planning. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3014–3026.
- [4] Zhao, W., Zhao, Y., Wang, S., & Qin, B. (2023). TransESC: Smoothing emotional support conversation via turn-level state transition. *Findings of the Association for Computational Linguistics: ACL 2023*, 6725–6739.
- [5] Welivita, A., Yeh, C.-H., & Pu, P. (2023). Empathetic response generation for distress support. *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 632–644.
- [6] Zhou, J., Chen, Z., Wang, B., & Huang, M. (2023). Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1714–1729.
- [7] Bao, Y., Zhao, H., Wang, X., & Liu, B. (2024). Multi-stream information fusion framework for emotional support conversation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 12032–12043.
- [8] Li, J., Peng, B., & Hsu, Y.-Y. (2024). Emstremo: Adapting emotional support response with enhanced emotion-strategy integrated selection. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 5794–5805.
- [9] Kang, D., Park, C., Kim, S., & Lee, J. (2024). Can large language models be good emotional supporters? Mitigating preference bias on emotional support conversation. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 15279–15295.
- [10] Feng, Y., Sun, X., Huang, L., Zhang, J., & Chen, Y. (2025). Effectiveness of AI-driven conversational agents in reducing depression among adolescents and young adults: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 27, e69639.